METHOD FOR ADAPTING A K-MEANS TEXT CLUSTERING TO EMERGING DATA

ABSTRACT

A method and structure for clustering documents in datasets which include clustering first documents and a first dataset to produce first document classes, creating centroid seeds based on the first document classes, and clustering second documents in a second dataset using the centroid seeds, wherein the first dataset and the second dataset are related. The clustering of the first documents in the first dataset forms a first dictionary of most common words in the first dataset and generates a first vector space model by counting, for each word in the first dictionary, a number of the first documents in which the word occurs, and clusters the first documents in the first dataset based on the first vector space model, and further generates a second vector space model by counting, for each word in the first dictionary, a number of the second documents in which the word occurs. Creation of the centroid seeds includes classifying second vector space model using the first document classes to produce a classified second vector space model and determining a mean of vectors in each class in the classified second vector space model, the mean includes the centroid seeds.